DOCUMENT RESUME

ED 14. 709 TM 006 532

AUTHOR	Hansen, Duncan N.; And Others
TITLE	Flexilevel Adaptive Testing Paradigm: Validation in Technical Training. Final Report May 1975-March 1977.
INSTITUTION	Air Force Human Resources Lab., Brooks AFB, Texas.
SPONS AGENCY	Air Force Human Resources Lab., Lowry AFB, Colo. Technical Training Div.
REPORT NO	AFHRL-TR-77-35-Vol-1
PUB DATE	Jul 77
CONTRACT	F41609-75-C-0040
NOTE	22p.
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS	Branching; Comparative Analysis; Complexity Level; Computer Programs; Cost Effectiveness; Individualized Programs; Military Personnel; *Models; Performance Factors; Post Secondary Education; *Technical Education; Testing; Test Reliability; *Test Validity; Time
IDENTIFIERS	*Adaptive Testing; Air Force; *Flexilevel Tests

ABSTRACT

A computerized adaptive testing model was assessed in a technical training system. The model, a modification of Lord's flexilevel paradigm, consisted of: the sequencing of test items in a difficulty hierarchy, adaptive entry of students into the test at a difficulty level appropriate to their predicted score, and systematic movement of students within the hierarchy based upon their performance on preceding items. Subjects were 444 airmen enrolled in an inventory management course employing individualized computer assisted instruction. Individualized entry was made according to reading test scores. The adaptive strategy was compared to conventional testing in terms of performance, test reliability, and test time. Both the adaptive and conventional test scores were obtained for each subject. The results revealed a high positive correlation between flexilevel and total test scores, and comparable reliability for both forms. The flexilevel test required fewer items than the entire test, yielding a length reduction of 39.5% and a time savings of 18.4%. For individualized entry, a slight time reduction was realized, without any noticeable change in performance, relative to the fixed entry control group. The benefits of adaptive testing included significant time reduction and maintenance of high standards of test validity. (Author/MV)

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

AFHRL-TR-77-35(I)

AIR FORCE HUMAN RESOURCES LABORATORY

ED143709

TM006 532

FLEXILEVEL ADAPTIVE TESTING PARADIGM: VALIDATION IN TECHNICAL TRAINING

By

Duncan N. Hansen
Steven Ross
Memphis State University
Memphis, Tennessee 38152

Dickie A. Harris, Capt, USAF

TECHNICAL TRAINING DIVISION
Lowry Air Force Base, Colorado 80230

July 1977
Final Report for Period May 1975 – March 1977

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Memphis State University, Memphis, Tennessee 38152, under contract F41609-75-C-0040, project 1121, with Technical Training Division, Air Force Human Resources Laboratory (AFSC), Lowry Air Force Base, Colorado 80230. Captain D. A. Harris, Instructional Technology Branch, was the contract monitor.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved for publication.

MARTY R. ROCKWAY, Technical Director
Technical Training Division

DAN D. FULGHAM, Colonel, USAF
Commander

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
AFHRL-TR-77-35(I)

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
FLEXILEVEL ADAPTIVE TESTING PARADIGM: VALIDATION IN TECHNICAL TRAINING

5. TYPE OF REPORT & PERIOD COVERED
Final
May 1975 – March 1977

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Duncan N. Hansen
Steven Ross
Dickie A. Harris

8. CONTRACT OR GRANT NUMBER(s)
F41609-75-C-0040

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Memphis State University
Memphis, Tennessee 38152

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
62703F
11210309

11. CONTROLLING OFFICE NAME AND ADDRESS
HQ Air Force Human Resources Laboratory (AFSC)
Brooks Air Force Base, Texas 78235

12. REPORT DATE
July 1977

13. NUMBER OF PAGES
20

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
Technical Training Division
Air Force Human Resources Laboratory
Lowry Air Force Base, Colorado 80230

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
adaptive testing
computer-assisted testing
computer-based adaptive testing
flexdevel adaptive testing
technical training

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
This study was designed to empirically assess a computerized adaptive testing model in an ongoing technical training system. The model was a modification of Lord's (1971) flexilevel paradigm and consisted of: (a) the sequencing of test items in a difficulty hierarchy, (b) adaptive entry of students into the test at a difficulty level appropriate to their predicted score, and (c) systematic movement of students up and down the hierarchy based upon their performance on preceding items. The subjects were 444 airmen enrolled in the Inventory Management/Materiel Facilities Course at Lowry Air Force Base, Colorado. They participated in the study as part of the normal achievement testing requirement for Block II of the course. Predictor variables for individualized entry were three reading tests administered to students prior to course enrollment. A within-subject design was employed

Item 20 (Continued)

in comparing the adaptive strategy to conventional testing on dependent variables or performance, test reliability, and test time. This involved first administering items according to the flexilevel algorith and then, after the student exited from the test at either the top or bottom end of the hierarchy, presenting all remaining items. Consequently, both adaptive and conventional test scores were obtained for each subject. The results revealed a high positive part-whole correlation between flexilevel and total test scores. Internal consistency indices for the two forms were essentially equivalent. With regard to length and time, however, the flexilevel test required nine less items than the entire test, yielding a length reduction of 39.5 percent with a concomitant time savings of 18.1 percent. Results for individualized entry were more indecisive as only a very small time reduction was realized, without any noticeable change in performance, relative to a fixed entry control group. The interpretation of findings stresses the potential benefits of adaptive testing in terms of significant time reductions and the maintenance of high standards of test validity.

TABLE OF CONTENTS

	Page
1.0 Introduction	1
2.0 Method	3
Subjects	3
Testing Materials	4
Computer Implementation	6
3.0 Results	6
4.0 Discussion	13
Reference	16

FLEXILEVEL ADAPTIVE TESTING PARADIGM: VALIDATION IN TECHNICAL TRAINING

1.0 Introduction

During the past decade technical training has initiated and expanded its commitment to individualized instruction. The dominant feature of individualization is the adaptation of instructional processes and resources to each student. Given the goal of adapting the overall technical training process, it seems only natural to ask to what degree can testing become adaptive? This study was an empirical assessment of the utility of the flexilevel adaptive testing paradigm (predictive entry and tailored item presentation for each student) within an ongoing Air Force technical training course. The primary purpose of the study was to assess the reliability and validity of embedded flexilevel adaptive tests by comparing the adaptive scores with scores on the conventional test (adaptive plus remainder).

Adaptive testing paradigms have grown out of the observation that many test items provide little or no information concerning training mastery, since they are either too hard or too easy for a given student. As a consequence of this observation, it seems only natural to find some appropriate way for removing those test items without detracting from either the reliability or the validity of the testing process. Numerous theoretical, simulated, and empirical investigations of adaptive testing (Hansen et al., 1974; Waters, 1975) have established both the framework and the scientific basis for adaptive testing. Unfortunately, two limitations have been observed. First, all of the empirical investigations of adaptive testing have tended to utilize ability measures; and in only one case (Ferguson, 1971) was the test content instructionally related. Secondly, in only a few cases (Larkin and Weiss, 1975; Waters, 1975) has the flexible resource of an interactive computer been utilized to further resolve the logistics of tailoring test item presentations for each student. Therefore, this study was an investigation of the generalizability of these prior adaptive testing findings to operational technical training under computer-based techniques.

In operational terms, the feasibility of adaptive testing for an ongoing technical training operation has yet to be established. Therefore, this study was implemented in an ongoing course at the Technical Training Center, Lowry Air Force Base, Colorado. Feasibility was to be judged in terms of the adaptation of students to the terminal, as well as the operational characteristics of the testing sessions. More importantly, did the adaptive scores yield direct equivalences to the total conventional scores so that they might be utilized for decision-making in training?

The preponderance of prior empirical investigations have utilized multiple groups or a concurrent validity approach. To establish the full implication of adaptive testing, a within-subject design was required to control for task difficulty and minimize the individual difference error variance. Further, the within-subject design fulfills course personnel's desire for the student to be given the total test. Related to this within-subject validity was an assessment of the impact of the adaptive algorithm in two procedural ways: the first, related to the prediction of the student's likely outcome score and entering him/her at an appropriate level within the computer-based adaptive test (adaptive entry); the second, the flexilevel algorithm developed by Lord (1971) which allowed a student to systematically move among harder and easier items according to a response contingency rule.

For technical training, cost implications, especially in the potential for reduced testing time, are critically important. It has been determined that a conventional test can be reduced in length by 40 to 50 percent (Hansen et al., 1974; Waters, 1975) as well as a somewhat equivalent reduction in total test time. The total test time was not reduced proportionately since it was found by Waters (1975) and Larkin and Weiss (1975) that adaptively presented test items required more mental processing time since they were on the "cutting edge" of the student's mastery level. It has been reported that the increase in testing time varies between 12 to 20 percent per item. Therefore, total test time was a major variable within this study.

From an operational training viewpoint, adaptive testing must demonstrate a number of advantageous outcomes if it is to be accepted. First, the adaptive and conventional scores should be essentially identical unless the adaptive process provides either new information or a significant reduction in error variance. This test score equivalence can be assessed by test correlations and the number of missed item after adaptive test cutoffs. These missed items could reflect critical training objectives and imply the need for remedial training. Second, the psychometric characteristics of both testing procedures should be essentially equivalent unless adaptive testing yields superior indices. A determination of reliability and validity indices can determine this issue. The amount of test time savings and required costs determine benefits. Finally, the implications for systems effectiveness were critical.

2.0 Method

The primary purpose of the study was to validate the adaptive flexilevel testing paradigm with respect to its major components of predictive test entry and tailored item presentation. Procedures for data collection involved the use of a repeated-measure design in which students were first entered into the test and then administered items by means of the flexilevel adaptive algorithm. After the student completed the adaptive portion of test, all remaining items were presented. Thus, both an adaptive score and a conventional test score were obtained for each subject in the sample.

Since scores on predictor variables for individualized entry were unavailable for the first 158 subjects tested, it was decided to enter these individuals at the median difficulty item in the test. This group was then treated as a unique treatment group for analyses of the effects of individualized versus standardized (median difficulty) test entry on dependent variables of: (a) test item, (b) item reduction, (c) reliability, and (d) validity of flexilevel performance scores. A within-subject design was also employed to assess the relationships between predictive entry and scores obtained on the adaptive and the conventional tests.

In summary, major independent variables consisted of: (a) the testing algorithm (adaptive versus conventional) with its final score, and (b) entry (individualized based on regression techniques versus median difficulty item). Dependent variables consisted of: (a) conventional test scores, (b) flexiscore, (c) number of flexi-test items, (d) flexi-time, (e) total test time, and (f) errors after flexi-exit. Reliability estimates of all test forms were obtained by means of the KR-20 procedure.

Subjects

The subjects consisted of 444 airmen enrolled in the Inventory Management/Materiel Facilities (IM/MF) course at Lowry Air Force Base, Denver, Colorado. When terminal equipment was available, the next student finishing the module under individualized training was selected. The student population from which the subjects were selected was considered as fairly homogeneous in characteristics pertaining to age, educational background, career goals, and military experience. Profile data collected during the past year indicated that the typical student enrolled in the IM/MF course was male (75 percent males to 25 percent females), an average age of 20, a high school graduate, and a relatively recent inductee into the Air Force (i.e., less than one year of experience).

Subjects were oriented to believe that participation in the study simply involved taking their regularly assigned achievement test (Block II) under a newly developed computer-assisted test administration system, that is, at an interactive computer terminal. Since the transition from adaptive to conventional item presentations took place without interruption or change in subjects became aware of the purposes of the experiment. It was doubtful that they even suspected that there was anything unusual about the selection or sequencing of items as compared to the conventional paper-and-pencil test of Block II.

Testing Materials

Predictor Variables. The measures employed as predictor variables for individualized entry were three reading tests normally administered to students prior to their formal admission to the IM/MF course.

The three tests were intended to provide estimates of individual aptitudes and abilities for comprehending and interpreting written information. Among the specific types of skills tested were general vocabulary, specific job sample, and reading test simulating the tasks associated with the Inventory Management career field. Descriptive test data derived from previous administrations to students in the IM/MF School (N = 367) showed means and standard deviations of 22.0558 and 7.0136; 4.7436 and 1.8629; and 6.0726 and 1.9232 for the three tests. Reliability estimates (KR-20) for the three test were .8573, .4295, and .5412, respectively.

Criterion Test. The Block II test of the IM/MF course was selected for use in validating the adaptive testing paradigm. This block covered researching supply publications and catalogs.

Satisfactory performance on the Block II test was considered prerequisite for progressing to more advanced concepts and skills taught in Blocks III and IV. The test consisted of 25 multiple-choice items, each containing four alternatives. Normative data collected during the past year indicated that mean student performance was 78.87 on a 100-point scale with a standard deviation of 13.22 points.

Procedure. Preparation activities involved meeting with course instructors and supervisory personnel from the IM/MF course two weeks prior to conducting the actual study to insure that the teaching staff understood the procedures that they would be required to follow in coordinating the test administration and data collection. Additionally, all instructors received a manual which provided a brief overview of the purposes of adaptive testing along with a detailed step-by-step account of the operational requirements for the present flexilevel test; that is, procedures for "signing on" the system, entering data, responding to items, interpreting and recording results, and "signing off."

The IM/MF course followed a criterion-referenced format in which schedules for program pacing and evaluation were largely self-determined by students. This necessitated administering the adaptive test on an individual basis and when participants elected to take Block 2 assessment. Specifically, once students informed the course instructor that they were ready to take the Block 2 test, they were directed to the terminal and given instructions for "signing on." Various panel displays then appeared in a prearranged sequence with progression from one display to another dependent upon the student keypunching appropriate symbols or words.

After "signing on," students entered identification information and scores obtained on the three reading aptitude tests. Students unfamiliar with the system were then given instructions for taking the computer test and for using the computer system in general. These instructions could be recalled any time questions arose during the actual test; also, students were encouraged to seek assistance from the laboratory instructor if ever uncertain about the proper procedures for responding.

Following preliminary instructions, students were entered into the flexilevel test at a difficulty level commensurate with their predicted performance (as determined by their reading aptitude scores). When such scores were unavailable, entry took place at the median difficulty item. Test items were administered separately with the rate of presentation determined entirely by the student. Procedures for responding simply involved keypunching the numbers of selected multiple-choice alternatives. Students were told to carefully consider their responses before continuing with the next item. If dissatisfied with their initial choice, they were to erase it and select another alternative; if satisfied, they were to finalize their answer by requesting that a new item be presented. Once answers were finalized, they could no longer be changed.

For the flexilevel portion of the test, the sequencing of items was determined in the following manner: once students were entered in the test at individually assigned levels, they were moved up and down the difficulty hierarchy (all items rank-ordered from easy to hard) based upon their performance. Specifically, each wrong response resulted in the presentation of the next easier unpresented item, whereas each correct response resulted exiting out of the hierarchy at either the top or bottom level, they were administered all remaining items. The test terminated after all 25 items were presented. At the completion of the entire test, the instructor was called to the terminal where he was able to obtain a summary of the student's performance. The specific information provided consisted of: (a) total test score (number correct times four), (b) individual item scores, and (c) total test time. For use by members of the research team the following were provided: (a) flexilevel score (proportion of correct items time 100), (b) flexilevel entry, (c) flexilevel exit, (d) flexilevel test time, and (e) a Green Score (average difficulty of correct items. A printed copy of the data was typically made available on the following day.

Computer Implementation.

Figure 1 presents a flow chart of a student moving through each of the steps. A more detailed description follows:

In signing on, the student entered his/her name and the computer executed a security check designed to limit system accessibility and assure test security. The system also determined the student's entry level in the test as he/she executed this test on the computer terminal.

When the student had completed the flexilevel portion of the test, the remaining test items were presented and the student's responses were evaluated (see Figure 1). The flexilevel portion of the test was evaluated in the post-analysis while the entire test was evaluated using standard Air Force performance criterion scoring procedures.

3.0 Results

During the early stages of the experiment, it was not possible to gain the entry predictive scores on the reading tests until the experiment had been running for approximately eight weeks. Therefore, two natural groups occurred: those in a fixed entry group (i.e., entered constantly at Item #13 within the item difficulty hierarchy); and a variable group which was entered according

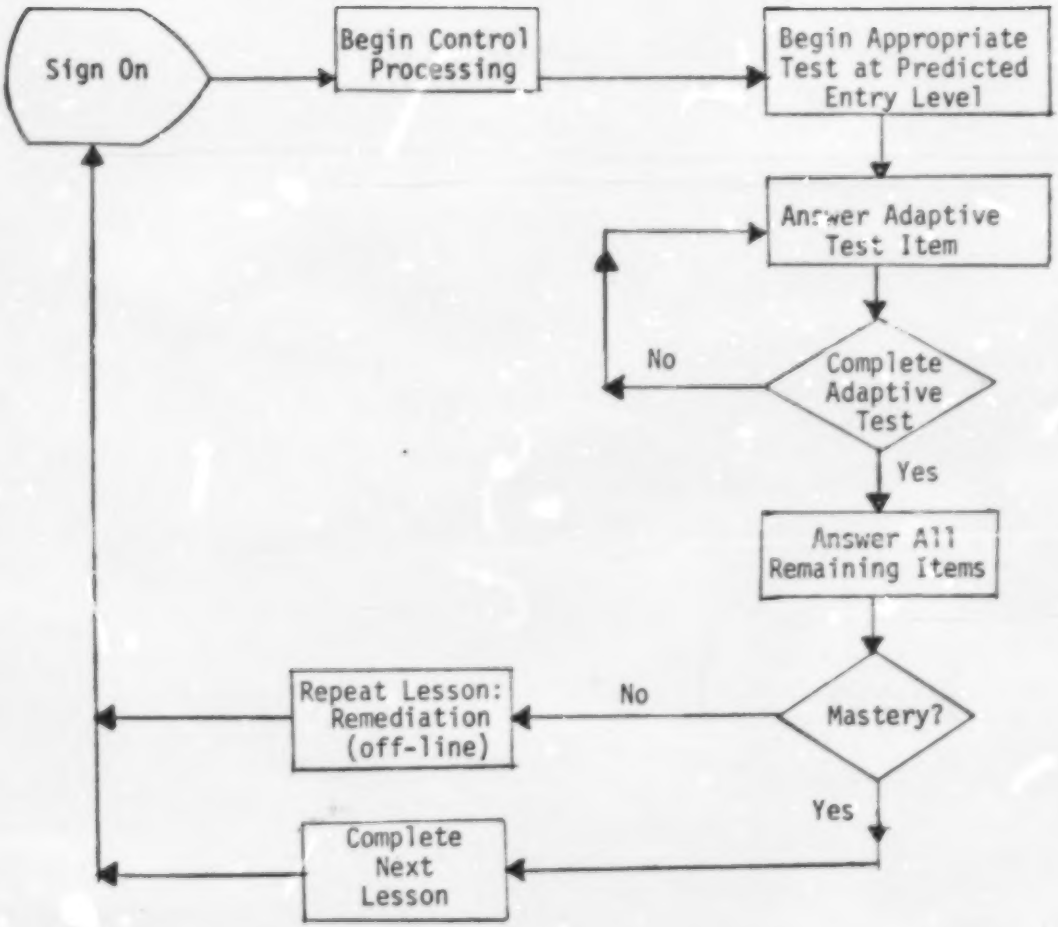

Figure 1. Flowchart of Student Progress Through Flexilevel Testing Program

to their expected outcome scores (i.e., the variable students were entered at the item that most closely approximated their estimated score for the total test). This procedural change allowed assessment of the impact of variable entry in comparison to a fixed one within the adaptive test paradigm (all known group characteristics were comparable).

Table 1 presents the means and standard deviations for each of the two groups on the critical variables. A comparison of the fixed and variable groups indicated that the means were essentially equivalent (all "t" values yield $p > .05$). The score distribution was almost symmetrical with limited skewness towards the criterion level of .70. Therefore, classical statistical methods can be applied to the data. Most importantly for the study, the flexilevel and total mean scores were practically equivalent ($p > .05$). The entry item number indicates mean level predicted for all the variable entry students, approximately one test item from the fixed level towards the more difficult items. Finally, the number of errors committed on the average by students after their flexilevel exit point is shown; one error tended to occur for these approximately 10 items (25 total items minus exit mean of 15.1). These errors are important to note in that if a flexilevel procedure were operational, it would not flag these items for the purpose of diagnosis and subsequent remedial prescriptions.

Table 2 presents the item statistics and reliabilities of the tests. The flexi-test was based on those students exiting after 15 or less items, but included all their responses. The total test referred to all students and items. As indicated, item difficulties tended to be relatively easy. This is a test characteristic commonly found in criterion-referenced testing. In turn, one can note that the reliability indices for items varied from a very low level up through a moderate level (the reliability index was derived by multiplying the correlation of the item score with the total score times the item's standard deviation). As presented in Table 2, the Kuder-Richardson 20 reliability coefficients were .594 for flexi students and .621 for the total group. The standard errors of measurement were 1.66 and 1.95, respectively. A comparison of biserial and phi coefficients, a contrast of classical and criterion-referenced methods, yield values similar in magnitude, the largest being .11. In reference to the initial 88 paper-and-pencil test protocols ($N = 88$) utilized to establish item difficulty and implement the adaptive testing approach, the Kuder-Richardson reliability was .586. Therefore, there was a slight positive increment in the internal consistency of the adaptive test presented over a computer terminal.

Table 1

Adaptive Test Descriptive Statistics
For Fixed and Variable Entry Groups

Variable	Fixed Entry (N = 158)		Variable Entry (N = 286)		"t" Value
N	$\bar{X}$	SD	$\bar{X}$	SD	
Total Score (25 Items)	77.7	11.1	77.2	11.1	.29
Flexi-Score	75.8	11.3	76.3	11.3	.32
Entry Item	13.0	NA	11.9	2.3	-
Exit Item	15.1	3.8	15.2	3.9	.53
Total Time (Min)	13.2	4.5	13.2	5.5	.28
Flexi-Time (Min)	10.3	4.6	10.3	5.1	.31
Errors After Cutoff	1.03	.96	1.01	1.0	.20

Table 2

Item Statistics and Reliabilities
For Flexi-Test and Total Test

	Flexi-Test (N = 289)		Total Test (N = 444)	
Items	Difficulty	Rel Index	Difficulty	Rel Index
1	.962	.039	.941	.045
2	.891	.048	.857	.077
3	.817	.095	.782	.155
4	.798	.140	.748	.205
5	.783	.174	.757	.152
6	.592	.216	.435	.160
7	.786	.133	.730	.097
8	.692	.174	.640	.119
9	.776	.203	.746	.157
10	.641	.226	.605	.167
11	.913	.036	.887	.070
12	.742	.127	.739	.107
13	.788	.138	.741	.119
14	.895	.141	.857	.118
15	.957	.010	.943	.058
16	.870	.117	.844	.135
17	.815	.149	.785	.147
18	.839	.167	.825	.133
19	.827	.138	.798	.127
20	.901	.068	.878	.094
21	.569	.167	.569	.180
22	.780	.162	.744	.163
23	.804	.159	.737	.177
24	.867	.098	.943	.062
25	.895	.143	.771	.145
KR-20 Rel.	.594		.621	
S.E. Meas.	1.663		1.951	

Analysis of flexi-test item sequences yielded reliability coefficients slightly lower in magnitude (15 items--r = .483; 18 items--r = .514; and 21 items--r = .573). These were derived by grouping at these exit points and item analyzing only the flexi-attempted items. The KR-20 indices were equivalent and non-significant ($p > .05$) especially if a test length correction was made.

In reference to the issue of variable test entry posed for the study (Table 1), the fixed and variable entry groups had similar means. There was no statistical difference ($p > .05$). More importantly, an assessment of the difference between the total score and the adaptive flexilevel score indicated no significant difference. Therefore, scores yielded under either approach could be used operationally within technical training.

In reference to the time measures, there were no tendencies for group differences. Most importantly, the flexilevel time was significantly less than the total time. This, of course, is to be expected since there were essentially nine fewer test items presented (see Table 1, Mean Exit Value). Average time per item was 31.68 seconds for the total test and 40.92 seconds per item for the flexilevel test, or an average of 29% longer per item on the average for the students to complete the adaptive item. It should be noted though, that the differences in time indicated that the remaining nine items were considerably easier and they also required far less mental processing time on the average. Both time measures (total and flexi) included a variable terminal orientation time of approximately three minutes.

Due to a computer recording error, item latencies were not collected. Although an item time average based on either the flexi-or-total times divided by the actual item numbers was not equivalent to item latencies, the total group yielded an average flexi-item time of .815 minutes and an average post-flexi-item time of .281 minutes. The post-flexi-items were obviously easier due to their .90+ level of difficulty and yielded shorter times by a factor of three. On the other hand, if one considers the eight percent of students who exited on the easy end of the item array and then took more difficult items, the average flexi-item times were .758 minutes and the average post-flexi-item times were .829. These post-flexi values were approximately equivalent to the total group's flexi-item time values.

In reference to the question concerning the functional relationship between adaptive test scores and total test scores, Table 3 presents the correlation among the significant variables. All coefficients were statistically significant. The part-whole

Table 3

Pearson Product-Moment Correlations
For Total Group

Variables	1	2	3	4
1. Total Score	-	.940*	.267	.306
2. Flexi-Score		-	.223	.281
3. Total Time			-	.680*
4. Flexi-Time				-

* Part-Whole correlation

correlation between adaptive flexilevel scores and total scores was R = .940. Note, there is a minimum built-in correlation, i.e., total score equals flexi-score plus the sum of other correct items divided by the number of items. A rank-order correlation procedure yielded a value of R = .902; thus the class rank position was highly stable between flexilevel and total score approaches. To maintain independence, the correlation of flexilevel scores with remaining item scores was r = .838. The reduction in the magnitude of the correlation can be attributed to the reduced variances on the remaining item scores. This correlation coefficient agreed with the near equivalent mean outcomes.

As presented in Table 3, total score had a low negative relationship with total test time; a similar negative relationship with flexilevel time (i.e., the higher the total test score, the shorter the flexilevel time); and a substantial negative relationship with the exit item number (i.e., the higher the total test score, the lower the exit item number). The adaptive test scores had a relational pattern which was highly similar to that of the total test scores; that is, a low negative relationship to total time, a similar negative relationship with flexi-time, and a substantial negative relationship to the adaptive item exit number. The relationship between total time, flexilevel time, and other variables was similarly patterned.

4.0 Discussion

The primary focus of this study was concerned with the operational validation of adaptive testing. The direct comparison of adaptive test scores with total test scores yielded a part-whole correlation coefficient of R = .940. The mean values and standard deviations for the two scores were practically equivalent. Viewed from this perspective, adaptive testing was a most appropriate substitute for a more conventional assessment approach in that it yielded highly equivalent scores having equivalent means and standard deviations. For the purposes of instructional decision-making, the two scores yielded identical outcomes.

Unlike prior studies, such as Waters (1975), the number of items reduced within the total criterion-referenced test was 39.5 percent as opposed to an expected value closer to 50 percent. Given that a student had to take a minimum of 12 items to exit from the flexi-routine, a saving of 70.9 percent was achieved on the remaining items. Since percentages are relativistic, the saving of 9.22 items out of 25 items was the important finding given the brevity of the test.

Even more important, as reported by Waters (1975), the reduction in testing time was only 18.4 percent and could be attributed to the increased mental processing time required for the relatively more difficult items presented under the flexilevel routine. For the total group the average flexi item time was .815 minutes while the average post-flexi-item time was three times less (.281 minutes). This finding tended to reverse for poorer students who exited at the easy item end (average flexi-item time = .758 minutes and average post-flexi-item-time = .829 minutes Additionally, it should be noted that approximately three minutes should be partitioned out for the time devoted to introducing a student to the computer terminal. The amount of time savings due to adaptive testing, especially for criterion-referenced tests, is likely to be considerably less than the proportional number of test items. A detailed item latency analysis should resolve this dissimilarity.

In a more limiting vein, the errors after the flexilevel cutoff were 1.01 errors out of 9.22 items. This value was within the range of the standard error of measurement for the total test but undoubtedly unacceptable for the assessment of specific training objectives. If the student population was divided by the exit ends of the test (hard vs. easy), one would find that the higher performers committed only .94 errors, while the lower performers committed 2.62. This would imply that for those students performing below the expected mean on an adaptive test, subsequent test items ought to be presented to more fully diagnose specific training objective accomplishment. As an attenuating factor on this mean error after adaptive testing cutoff, the item difficulties utilized within the study had a Spearman rank-order correlation of r = .68 with the difficulties generated by the 88 standardizing students. This fluctuating item difficulty was probably due to both the shifting nature of instruction within the course as well as the possibility of shifting student abilities. The major point was that adaptive testing would require constant monitoring of the item difficulty hierarchy to be effectively pursued. Given the shifts in item difficulties as presented in Table 2, adaptive testing was robust in that the adaptive scores and total scores were quite similar. This suggested that initial implementation procedures were not likely to yield spurious results. In considering all of the above, it seems appropriate to consider adaptive testing an acceptable alternative to conventional test item presentation.

In reference to the fixed versus variable entry, the results were far more indecisive. Mean values were nearly equivalent.

Given the logistic requirement of assembling entry predictor measures and utilizing these in a regression approach, the implementation of flexible entry is questionable. If one had entered all of the students at the expected mean value for the test, one could anticipate that a nearly optimal result would be forthcoming. On the other hand, if the number of test items was vastly increased (N =50) and with even greater ranges of item difficulties, flexible entry might prove to be of greater benefit.

In reference to the time savings gained through adaptive testing, it should be noted that conventional approaches would have allowed each student 30 minutes to complete the 25 items. If this was considered the benchmark, one would have anticipated a savings of 59.6 percent as opposed to the 18.4 percent noted. In addition, the time required for the computer terminal directions could be further reduced if terminal-oriented testing became a comprehensive part of the training operation. The most impressive aspect of computer-based adaptive test time was the obvious reduction in comparison to a conventional test. There were anecdotal reports from students concerning the stressful aspects of computer presentation. There were complaints specifically directed at the inability to alter answers after initial entry. Further, the computer stressing effects undoubtedly accounted for the slightly higher reliabilities found within adaptive testing, a finding well-documented by Hedl (1971).

The feasibility and validity of adaptive testing as an integral part of computer-based technical training has been documented by this study. The high validity and reliability indices supported this finding. The comparison of the fixed and variable groups also contrasted the first and second eight-week periods. There were no discernible differences. The fluctuating item difficulties did not yield negative impacts. The reported operation was smooth with some initial criticism. Therefore, the system implications of adaptive testing were positive.

This study established both the positive aspects as well as the limitation of adaptive testing within ongoing technical training. Further efforts are planned to study the predictive validity found in a multiple or hierarchically arranged test paradigm. While many obvious extensions in the reliability and validity areas remain for further study, it is clear that the essential outcomes in this study as well as the Waters (1975) and Larkin and Weiss (1975) studies indicated its obvious advantage for both improved assessment and reduced training time within a technical training system.

References

Ferguson, R.L. Computer Assistance for Individualized Measurement. Report, 1971-8, University of Pittsburgh, Learning and Research Development Center, 1971.

Hansen, D., Johnson, B., Fagan, R., Tam, P., & Dick, W. Computer-based adaptive testing models for Air Force technical training environment phase I: Development of computerized measurement systems for Air Force technical training. AFHRL-TR-74-48, AD-785-142. Air Force Human Resources Laboratory, Technical Training Division, Lowry AFB, CO, July 1974.

Hedl, J.J. An Evaluation of a Computer-Based Intelligence Test. Technical Report, No. 21, Florida State University, Tallahassee, 1971.

Larkin, K.C., & Weiss, D.J. An Empirical Comparison of Two-Stage and Pyramidal Adaptive Ability Testing. Research Report, 75-1, University of Minnesota, Minneapolis, 1975.

Lord, F.M. A Theoretical Study of the Measurement Effectiveness of Flexilevel Tests. Educational and Psychological Measurement, 31 (Winter 1971): 805-813.

Waters, B.K. Empirical investigation of the stradaptive testing model for the measurement of human ability. AFHRL-TR-75-27, AD-A018-611. Williams AFB, AZ: Flying Training Division, Air Force Human Resources Laboratory, October 1975.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1977 – 771-067/48